DOCUMENT RESUME

ED 070 770 TM 001 639

AUTHOR	Fowler, Ernest P.; Bramble, William J.
TITLE	An Analysis of Personality Data Using Rasch Measurement Model.
PUB DATE	[71]
NOTE	21p.; Working draft
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Data Analysis; Factor Analysis; *Goodness of Fit; Grade 7; Grade 11; Grade 12; *Mathematical Models; *Measurement Instruments; *Personality Tests; Questionnaires; Statistical Analysis; Test Results; Tests
IDENTIFIERS	High School Personality Questionnaire; *Rasch Measurement Model

ABSTRACT

The applicability of the Rasch model to data from a typical personality test, the High School Personality Questionnaire (HSPQ), was studied. The data were gathered on Junior High and High School students in the Louisville Public Schools (Kentucky). Item easinesses and person abilities were estimated and compared by age group, within each age group, and across two points in time for the older age group. In addition, certain results from the Rasch analyses were compared with those of factor analysis. A sample of 1,000 students was taken from each of the groups (Junior High, 7th graders; Senior High, 11th and 12th graders). Results of the study are related to five questions considered. The first question was whether or not there were patterns of fit to the Rasch model when responses are dichotomized in different ways. The results indicated that no single key was superior to others in producing fit. The second question was concerned with fit of the model for the data considered; it was found that frequently there was lack of fit, but it is noted that the test statistic was conservative. The third question related to the stability of item easinesses estimates within a group and across two points in time for that group. The conclusion was that different item easinesses are obtained when different degrees of possession of the trait are focused upon. The fourth question was how stable the tests of fits results are across time; pre- and post-comparisons of fit found 55% in agreement. The fifth question concerned how the item mean squares are related to factor loadings; in almost all cases, the item with the highest mean square was also the item with the lowest loading. (Author/DB)

ED 070770

5.10

AN ANALYSIS OF PERSONALITY DATA USING RASCH MEASUREMENT MODEL [1]

Ernest P. Fowler and William J. Bramble

University of Kentucky

TM 001 639

[1] This paper is a working draft and should not be quoted or cited without the authors' permission.

Introduction

The Rasch measurement model is a mathematical statement of the odds of a correct response when a person of certain ability encounters an item of given easinesses. The attractive property of this model is that item easinesses and person abilities can be estimated independently. Thus, the term "sample-free" has been applied to the model. The model was conceived (Rasch, 1960) for application to ability data. Thus most treatments of the model speak of binary (correct vs. incorrect) responses to items. However, there has been little application of the model to personality test data which, typically, is less reliable than ability test results.

The purpose of this paper is to study the applicability of the Rasch model to data from a typical personality test, the High School Personality Questionaire (HSPQ). The data were gathered on Junior High and High School students in the Louisville Public Schools. Item easinesses and person abilities (these terms are defined below) were estimated and compared: by age group (Junior High students and Senior High students; within each age group (cross-validation); and across two points in time (beginning vs. end of the 1970-71 school year) for the older age group in order to look at the stability of the estimates. In addition, certain results from the Rasch analyses were compared with those of factor analysis.

Before discussing specific procedures of this paper, however, it will be useful to review some important aspects of the Rasch model.

The responses to ability test items are usually scored as correct (+) or incorrect (-). Personality tests, however, typically have item

alternates which are weighted in varying degrees, such that choice of an alternate with a larger weight results in a larger score on a particular dimension for the person. Thus, there is no "correct" alternate, but alternates of different strengths with respect to the personality trait measured. One solution to the problem of non-dichotomous scores on a personality test is to force a binary scoring routine. This procedure was used in several forms in this paper. The results are described below. Though a formulation of the model which permits polychotomous responses would be the optimal analytic tool here, it has not been fully developed to date. The results presented here may be of some interest to persons working on the polychotomous form of the model.

Basic Assumptions of the Model

Scoring the responses as + or - for person ν on item i of a unifactor test allows us to represent a correct response as $a_{\nu i} = +$. The probability of a correct response, then, is stated as $p\{+|\nu, i\}$.

The first assumption of the Rasch measurement model is given as

$$p\{+ \mid \nu, i\} = \lambda_{\nu i}/(1 + \lambda_{\nu i}); \lambda_{\nu i} \geq 0.$$

Here $\lambda_{\nu i}$ is the odds of success on item i for person ν.

The second assumption of the model is that

$$\lambda_{\nu i} = \xi_\nu \varepsilon_i; \quad \xi_\nu \geq 0 \text{ and } \varepsilon_i \geq 0$$

where ξ_ν is the ability of person ν and ε_i is the easiness of item i. This equation states that the odds of person ν responding correctly to item i are the product of the person ability (ξ_ν) and item easiness (ε_i) components.

Let $a_{\nu i} = 0$ if person ν responds incorrectly to item i and $a_{\nu i} = 1$ if person ν responds correctly to item i. Then, given the above assump-

tions, the probability of a correct response for person v and item i is given as

$$p\{a_{vi} \mid v,i\} = (\xi_v \varepsilon_i)^{a_{vi}} / (1 + \xi_v \varepsilon_i).$$

The third assumption of the model is that all answers, given the parameters, are stochastically independent.

Estimates of the Parameters of the Model

There are different procedures (see Rasch, 1960; Bramble, 1970; Wright and Panchapakesan, 1968) for obtaining estimates and standard errors of person abilities and item easinesses.

The maximum likelihood estimation procedures (Wright and Panchakesan, 1968) involve obtaining iteratively a solution to the implicit equations

$$a_{+i} = \sum_{i}^{k-1} (r_i \exp(b_j^* + d_i^*)/(1 + \exp(b_j^* + d_i^*))), \quad i = 1,2 \ldots k$$

$$j = \sum_{i}^{k} (\exp(b_j^* + d_i^*)/(1 + \exp(b_j^* + d_i^*))), \quad j = 1, 2 \ldots k-1$$

where a_{+i} = number of persons who get item i correct (item score)

j = the score, an ability estimate is obtained for each score

r_i = number of persons in score group j,

b_j^* = ability estimate

d_i^* = easiness estimate

The Newton-Raphson procedure is used to solve for the unknown parameter estimates.

An approximation to the standard variance of item estimates is given as:

$$V(d_i^*) \simeq 1/ \sum_{j}^{k-1} (r_i \exp (b_j^* + d_i^*)/(1 + \exp (b_j^* + d_i^*))^2).$$

An approximation to the standard variance of ability estimates is given as:

$$V^* (b^*) = 1/(C(b^*) \exp (b^*)) + (1/C^2(b^*)) \sum_{i}^{k} (V(d_i)(\exp (d_i)/(1 + \exp (d_i + b^*))^2)^2)$$

where

$$C(b^*) = \sum_{i}^{k} (\exp (d_i)/(1 + \exp (b^* + d_i))^2).$$

The fit of an item to the model may be investigated by forming standard deviates of the score group X items matrix,

$$y_{ij} = (a_{+i}^{(r)} - E (a_{+i}^{(r)})) / \nu (a_{+i}^{(r)}))^{\frac{1}{2}}$$

where

$$E (a_{+i}^{(r)}) = rh^{(r)}$$

and

$$\nu (a_{+i}^{(r)}) = rh^{(r)} (1 - h^{(r)}).$$

The mean square for a particular item may then be the criterion by which one can determine fit (small mean square) or lack of fit (large mean square).

The overall test of fit is made using a chi-square X^2 statistic, a conservative test. It is obtained by summing all squared unit normal deviates,

$$X^2 = \sum_{1}^{r} \sum_{1}^{r-1} y_{ij}$$

with df = (score groups - 1)(items - 1). A likelihood ratio statistic

may also be used as a test for fit.

We now turn to the basic questions posed by this paper. They may be summarized as follows:

1. Are there patterns of fit to the Rasch model when responses are dichotomized in different ways?

   For the data considered in this paper, the question is: should all responses which indicate a positive amount of the trait, only responses suggesting an extreme possession of the trait, or responses indicating a moderate level on the trait be scored as "correct"?

2. How well does the Rasch model fit for the data considered? Is there generally a greater degree of fit for one age group? Does a particular scoring key result in greater fit?

3. How stable are the estimates of item easinesses both within a group and across two points in time for that group?

Method

The Sample

In an experimental project involving public schools in the Louisville Public Schools, the HSPQ was administered to approximately 6,000 students, selected from grades one through twelve. The pretest data were obtained during September, 1970 and postest data were obtained during May, 1971. For this study, the data were obtained from two general groups: Junior high students (seventh graders) and Senior High students (eleventh and twelfth graders). A sample of 1000 students was taken from each group. So that cross-validations could be performed within each group, these samples were arbitrarily split into two samples of 500 each. In the paper, we shall refer to the first and second samples of the Junior High group and the first and second samples of the Senior High group though "first" and "second" are merely labels.

Instrumentation

The personality test used in the study was the IPAT HSPQ. This instrument has 14 subtests, each of which measures a different trait. One subtest (B), having one correct answer for an item, is an ability test which, it's authors report, measures "crystallized" as opposed to "fluid" general ability. The former (crystallized) kind of ability is thought to change over time and broadly refers to one's gradual acquisition of information (e.g., vocabulary). The latter type of ability (fluid) refers to abilities thought to be more stable, such as those required to see analogies or deduce a certain conclusion from given information. There is occasion, at a later point, to

describe two other subtests. Though the remaining eleven subtests are not described, their titles are reported in Table 1. These scales are designed to measure various other aspects of personality. Unlike subtest B, these scales contain items having alternatives representing three levels of response rather than having alternatives which are merely "right" and "wrong". Thus the items are trichotomous but the alternatives are assumed to be ordered in terms of strength.

Each subtest is comprised of ten items, each item having three alternates. The HSPQ scoring system is such that the weights of each alternate in its contribution to the total subtest score is either zero, one, or two. The weights for the single alternates per items that a person chooses, then, are summed to give his score. For a ten-item subtest, therefore, the highest possible score is 20.

Because of the necessity for dichotomous data, the 0, 1, or 2 responses were transformed to zero-one data. This was accomplished using the following keys: key-1 accepted only the alternates weighted one (middle alternatives) as "correct"; key-2 accepted only the alternates weighted two (extreme alternatives) as "correct", and key-1,2 accepted either the first or the second kind of alternate (any positive response = 1). To clarify, the responses weighted two represent alternates that characterize more strongly the trait measured; responses weighted one characterize with less strength the trait measured. For comparison, each of these keys was used in scoring each of the 14 sub-tests.

Results and Discussion

Table 1 shows probabilities of fit, first for the X^2 test and second for the likelihood ratio test for all samples and subtests. Before looking at particular subtests, some general trends should be noted. Though measures of skewness were not obtained, the general tendency observed was for key-1 to produce a positively skewed distribution of scores, key-2 to produce a distribution less positively skewed (or approximately normal), and for key-1,2 to produce a negatively skewed distribution. Recalling that a score group refers to all persons who obtained a certain score and that there are nine possible score groups for each subtest here (score groups zero and ten are discarded because they do not contribute to the analysis), it was observed that key-1 analysis sometimes resulted in as many as three empty score groups. The empty groups for this key were always at the upper end (i.e., either nine, eight and nine, or seven, eight, and nine). To be more specific about skewness, 80 to 90 per cent of the responses typically were contained in either score groups one through five or five through nine. Key-2 generally produced better fit to the model, though there are variations on this conclusion for individual subtests. Notice also that subtest B, the ability test, produced consistent lack of fit statistics. These distributions were generally negatively skewed (i.e., the items were easy).

In order to obtain information relevant to questions one and two (regarding results with different keys), the total number of subtest fits are shown in the margins of Table 2. The right margin shows total fits by key for each subtest. Analyses for subtests D and Q_3 generally resulted in better fit for all keys. Analyses of subtests H and I re-

sulted in a smaller number of fits.

The lower margin of Table 1 shows the total numbers of fits by key for each sample. The total number of fits for the Senior High group (48) is similar to the total for the Junior High group (51).

The totals of the marginal sums are shown in the lower right corner of Table 1. The scoring procedure keying on the middle response (key-2) shows a slightly larger number of non-significant statistics. When marginal values are inspected, however, it is seen that this key does not consistently produce more non-significant statistics either across samples (lower marginals) or across subtests (right marginals). Thus, the results do not clearly favor the use of one key over another.

After these preliminary analyses, certain subtests were selected for further comparisons, relating to question three, stated initially. Subtest D ("Excitability") was chosen because the results typically displayed consistent fit for all keys. Subtest I (Premsia) was selected for the opposite reason. Subtest B, the ability test, served as a kind of baseline against which D and I were compared. The psychological meaning of each subtest is extensively discussed in Handbook for the HSPQ (p. 26). Here, we shall briefly describe the nature of subtests D and I. The nature of subtest B has been discussed.

Subtest D measures one's "excitability" or tendency toward "irrepresible impulsiveness". The high scorer on the D scale is one who "reports that he is a restless sleeper, easily distracted from work by noise or intrinsic difficulty, is hurt and angry if not given important positions...." (Handbook for the HSPQ, p. 29) The low scorer reports that he is shy and prefers one or two friends rather than crowds. The

person who scores high on subtest I "shows a definite fastidious dislike of 'crude' people and rough occupations, a recreational taste for 'romantic travel' and new experiences...." (p. 31) The low scorer is described as one who is "tough-minded, rejects illusions."

In order to illustrate the item difficulty orders both for different keys in a sample and for the same keys across samples (i.e., cross validation), Table 2 was constructed. In each column of Table 2, the first number refers to the particular item, the second to the log easiness obtained for that item. Log easinesses are ordered from positive to negative values. A large negative value means that the item is "difficult" or, in this case, that few people answered it in a way the key would consider as "correct". We can then compare the positions of various items from one type of analysis to another (i.e., by key), or from one sample to another. Kuder-Richardson formula 20 (KR-20) reliabilities are also reported for each analysis.

First, let us look only at item rankings, comprising the first columns of Table 2. A general conclusion one can make about the rankings within a sample resulting from use of the three different keys is that there is no consistent agreement of item rankings. This is true for all selected subtests. In fact, the rank orderings for the different keys are typically slightly inversely related. In other words, quite different easiness estimates are obtained when one scores different strengths of a trait as "correct".

We now wish to investigate the stability of easiness orders resulting with a single key across different sets of data. Thus, we first look at orders for all samples for key-1. The same is done for key-2 and key-1,2. A general conclusion one can make upon such comparisons is that there is

moderate agreement from one sample to another, for either subtests D or I. There are exceptions. For example, it is seen that there is one case in which agreement from one sample to the other is quite high. This is in the case of subtest I, key-2 for the Junior High group; only the items positioned four and five are inverted.

The other case where there is considerable agreement of item orders from the First Sample to the Second Sample is for key-1,2 in subtest D of the Senior High Sample. Different items change rank in this case, however. The statistic for fit for these two analyses was .008 and .211. Interestingly, Subtest I resulted generally in more consistent item orders, by key, than did Subtest D, the better fitting subtest.

Subtest B for the Senior High Samples had only one inversion, items nine and seven. There was less agreement of orders for a particular key for the Junior High Samples. Agreement across samples within the Senior High group, however, was somewhat better. Note that items five and seven through ten did remain generally on the difficult end of the groups. The consistency, by key, from the Junior High to the Senior High Sample, for subtest D and I, was generally no better than that achieved within either group. And, no particular key gives better results than the others.

Now let us look at the easiness log estimates themselves, comprising the second columns of Table 2. Only easiness estimates for selected analyses will be discussed here. The reader may investigate for himself the remaining cases. Observe the easiness estimates for key-2, first and second samples of the Junior High group. Notice that items four and five inverted in rank from the first to the second samples. Note also, however, that the easiness estimates for these

items were similar for the first sample, becoming more alike in the second sample. Remaining estimates are quite agreeable.

In the other case for which there was only one inversion of rank (subtest D, key-1,2, Senior High group), the same tendency regarding similarity of easiness estimates existed. That is, items nine and six were fairly close in easiness on the first sample and also on the second, the difference being that their positions were reversed in the second sample. In other words, for the two examples related here, one might explain the reversals in easiness order by reference to the closeness of easinesses, or that their standard errors will include the other easiness estimate.

Finally, with respect to Table 2, Kuder-Richardson reliability coefficients (KR-20) obtained for each subtest, by key, are shown. For the poorly-fitting Subtest I, key-2 obtains higher coefficients with just one exception. Of the KR-20's for subtests D and I, .535 was the highest, with probability of fit being less than .001. The KR-20's for subtest B are not particularly higher, though they do tend to exceed the others slightly.

We now report analyses regarding question four, related to the Rasch analysis over time. First the postest data analyses are presented and discussed. Then, pretest results are compared with these, in terms of fit statistics and item easiness orders.

We shall use the postest data for only the Senior High group. This group contained the samples which tended to produce more consistent orders of easiness estimates.

Table 3 shows probabilities for the tests of fit. The format is the same as for Table 1. It can be seen that no particular key results in much better fit than another, results which are similar

to those of the pretest data. When we return to Table 1 and tally the number of fitting subtests by key, for just the Senior High group, the results are 12, 18, and 18 for the respective keys. Here, however, the first key produced a slightly larger number of non-significant tests. Again, as with pretest data analyses reported in Table 1, subtest D obtains in overall better fit and Subtest B results show consistent lack of fit, as does Subtest I.

To shed light on the question regarding consistency of fit to the model from one point in time to another, Table 4 was constructed. Non-fitting subtests are indicated by "L" and fitting subtests simply by "F", from pre to post in each case. No single key gives more consistent results. Subtest Q_3 gives the most consistent results across all keys from pre- to postests for either sample. Subtests D, O, and Q_2 show only moderate consistency, especially when key-1,2 is used for the latter two tests.

The second part of question four related to the consistency of item easiness orders over two points in time. Subtests B, D, and I again were selected so that postest results can be compared with those already reported. Table 5 contains the relevant information, in the same format as Table 2. First, we note characteristics within the postest data, then we compare easiness estimates from pre- to post-data.

First, let us look at the columns which contain the ranks of the easiness estimates. The rank agreement for the different keys within each sample is usually lacking or inverse. However, agreement across samples by key is high in three cases. One case involves subtest B, in which the positions of items four and two and items nine and seven

are inverted from the first to the second sample. The other two cases relate to key-1,2 of subtests D and I. While this key produced moderate correspondence of item orders for subtest I, high agreement is seen across samples for the D. subtest. These results are much like those reported in Table 2, which contains pretest analyses.

The item easiness rank correspondence from pretest to postest is somewhat higher. This is especially true for the second sample where key-2 and key-1,2 result in high correspondence for subtests D and I. For the single key used with subtest B, it is seen that item easiness rankings are rather consistent across all Senior High samples, with items two and four and seven and nine having some tendency toward inversion. The item orders of the Junior High samples agree more with themselves than with either pretest or postest samples within the Senior High group. At the same time, it is recalled that easiness orders are more stable within the latter group than the former.

The KR-20's obtained in the postest analysis of subtest D are consistently higher by about .10 for subtest D. For subtest I, key-1 of the first sample produced a sizeable increase (.266 to .436), otherwise no discernable pattern.

The final analysis is related to question five. Here, we want to investigate the relationship between mean squares obtained for each item using the Rasch model and loadings of each item resulting from a factor analysis.

A requirement of the Rasch model is that the test have a factor structure that is unitary. One aspect of the factor analytic method referred to as "unrestricted maximum-likelihood factor analysis" (Jöreskog, 1967) is that one may test an hypothesis that the data can be accounted for by a certain number of factors. In this analysis,

therefore, this factor analytic technique will be used to test the hypothesis ($\alpha > .01$) that a unifactor solution can be obtained for each null of the selected subtests. We shall observe in some detail the results for subtest B. We will then look at the overall outcome of the analyses of subtests D and I, then compare mean squares and loadings in a separate table.

Some cautions regarding the maximum likelihood method should be mentioned. The method should be used when the distribution of the variables is multivariate normal. For variables composed of responses to a single item, this is not a valid assumption. Choosing between this procedure or classical procedures, however, we decided to use the former in factoring the correlation matrix.

Table 6 contains, for each subtest, the factor loadings for each item for a single factor solution for each key. The X^2 probability for a two-factor solution is also given, though loadings on the two factors are not listed.

Look at the results in Table 6 for subtest B. We have seen that the Rasch model consistently does not fit for these data, though item rankings and easiness estimates across the High School samples are frequently more consistent than for other, better fitting subtests. This table shows that the first sample did produce a unifactor test, but that the second sample did not. Two items, five and eight, had near zero loadings on both analyses. (Recall that it was items seven and nine which were inverting orders from one sample to another in the Rasch analysis.) If we now return to the Rasch output and check the mean squares of the items for these samples, it is seen that while most items have mean squares at acceptable levels (e.g., slightly

above 1.0), the statistics for these items are quite large, being nine and twenty three, respectively, for the first sample and seventeen and twenty for the second sample. Checking the percentages of correct responses to these items shows they are most difficult, falling roughly at .35 and .22 for both samples. The authors rechecked the subtest key constructed by IPAT and found no errors. A short description of these two items may be useful.

Item five is an analogy item, paraphrasing: part is to half as parent is to _____. The alternates are (correct one is underlined) "grandfather," "father," and "son." Item eight requires a deduction, again paraphrasing: given five coins with three of them bent and four of them silver, how many silver coins must be bent? Alternate responses are: one, two, or three.

There is no reason to suspect that the content of these items differs greatly from the other items of subtest B. However, it is likely that these items are too difficult for this group.

The results for subtest B tend to show more unifactor solutions than the results for subtest I. No clear, overall trend in the X^2 statistics, by key, is apparent.

Table 7 shows, for each subtest, the items (first column) ranked according to mean square and the same set of items (second column) ranked by factor loading. Mean squares are ranked from low (top) to high (bottom) but loadings from high (top) to low (bottom). Thus, the first row contains the item (1) with the lowest mean square and the item (3) with the highest loading.

Another question of concern in this paper is related to a comparison of items based on the mean square fits obtained in the Rasch procedure and the loadings resulting from the Factor analyses. Table

7 contains items ordered according to mean squares and loadings for both Senior High samples, for the single key in the case of subtest B and for three keys in the case of subtests D and I.

It can be seen in Table 7 that in 11 of the 14 pairs of analyses the item with the largest mean square corresponds to the item with the smallest loading. From this extreme, the number of items that correspond tend to decrease and become further apart in rank, so that for most of the items there tends to be no particular pattern apparent. No particular key results in an order any more consistent across samples than another. Nor do different keys within a sample give similar orderings. Although it may not be apparent at first, closer inspection of Table 7 will show that subtest D tends to produce orders of mean squares and loadings that are closer together.

The results in Table 7 suggests rather strongly that items with large mean squares tend also to be the items that have small loadings on the factor. But this trend is not discernable as the items become closer in mean square value and in size of factor loading.

Summary

Though the Rasch model was conceived for application to ability tests, the authors undertook to apply the procedure to a personality test, the HSPQ. Though a polychotomous model would have been more appropriate for these kinds of data, the unavailability of such a model required the authors to consider different scoring procedures in order to make use of the model for dichotomous responses.

There were five questions considered. The first related to whether or not there were patterns of fit to the Rasch model when responses are dichotomized in different ways. The results indicated that no single key was superior to others in producing fit.

The second question was concerned with fit of the model for the data considered. Frequently, there was lack of fit, though it should be noted that the test statistic was a conservative one. For the pretest data roughly 66% of the analyses resulted in fit. For the postest data, approximately 54% of the analyses produced fit to the model.

The third question related to the stability of item easiness estimates within a group and across two points in time for that group. The conclusion was rather clear that different item easinesses are obtained when different degrees of possession of the trait are focused upon. For the same key, however, easiness estimates are somewhat consistent, moreso with older children.

Fourth, how stable are the results from the tests of fits across time? In the pre- to post-comparisons of fit, 55% were in agreement.

Finally, how are the item mean squares related to factor loadings? In almost all cases the item with the highest mean square was also the item with the lowest loading. As items became more similar in values on mean squares and on loadings, no relationship was apparent.

REFERENCES

Bramble, William J. A least square method of parameter estimation for the logistic measurement model. AERA presession, 1971.

Keesling, James Ward. Computer programming of the model. Presentation at the 1969 AERA Presession on Person-Free Item Calibration and Item-Free Person Measurement. Los Angeles, California.

Lord, Frederick M. and Novick, Melvin R. Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley, 1968.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research, 1960.

Wright, B. and Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 29, 23-48.